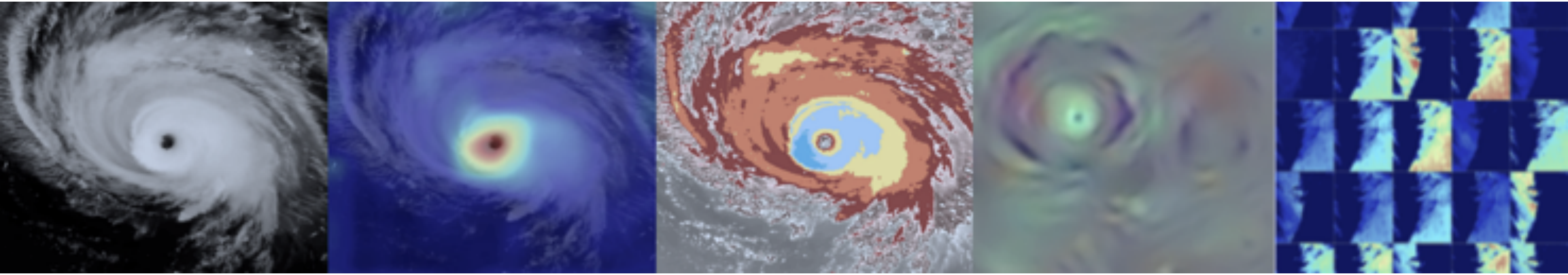
Advancing AI for Earth Science: A Data Systems Perspective



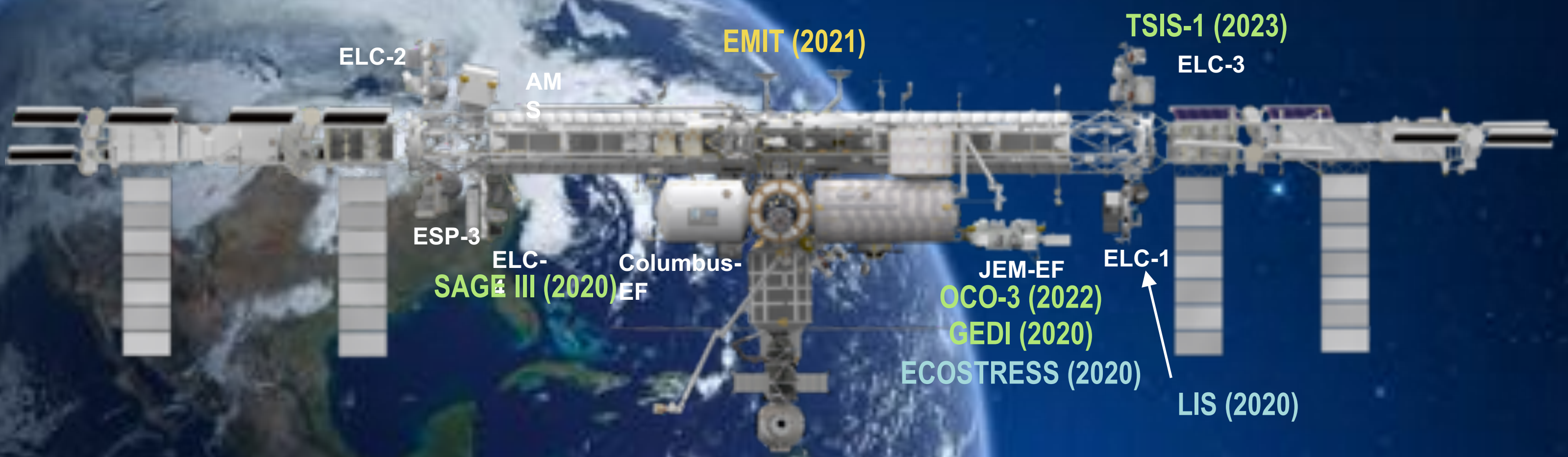
Manil Maskey, Ph.D.
NASA
Earth Science Data Systems/NASA Headquarters
Interagency Implementation and Advanced Concepts Team (IMPACT)/NASA Marshall Space Flight Center

ESA EO Φ-week 2020


NASA EARTH FLEET
OPERATING & FUTURE THROUGH 2023
SWOT (CNES)
LANDSAT-9 (USGS)
SENTINEL-6 Michael Freilich/B (ESA)
TROPICS (6)
GEOCARB
NISAR (ISRO)
MAIA
TSIS-2
TEMPO
PREFIRE (2)
PACE (NSO)
GLIMR
ICESAT-2
GRACE-FO (2) (GFZ)
CYGNSS (8)
NISTAR, EPIC (DSCOVR/NOAA)
CLOUDSAT (CSA)
TERRA (JAXA, CSA)
AQUA (JAXA, AEB)
AURA (NSO, FMI, UKSA)
CALIPSO (CNES)
GPM (JAXA)
LANDSAT 7 (USGS)
LANDSAT 8 (USGS)
OCO-2
SMAP
SUOMI NPP (NOAA) (JAXA)
INVEST/CUBESATS
RainCube
CSIM-FD
HARP
TEMPEST-D
CIRiS
CTIM
HyTI
SNoOPI
NACHOS
ISS INSTRUMENTS
EMIT
CLARREO-PF
GEDI
SAGE III
OCO-3
TSIS-1
ECOSTRESS
LIS
JPSS-2, 3 & 4 INSTRUMENTS
OMPS-Limb
LIBERA
6.5.20
(PRE) FORMULATION
IMPLEMENTATON
PRIMARY OPS
EXTENDED OPS

INTERNATIONAL SPACE STATION

EARTH SCIENCE OPERATING MISSIONS



EXPRESS Logistics Carriers: ELC-1, ELC-2, ELC-3
External Stowage Platforms: ESP-3
Alpha Magnetic Spectrometer: AMS
Columbus External Payload Facility: Columbus-EF
Kibo External Payload Facility: JEM-EF

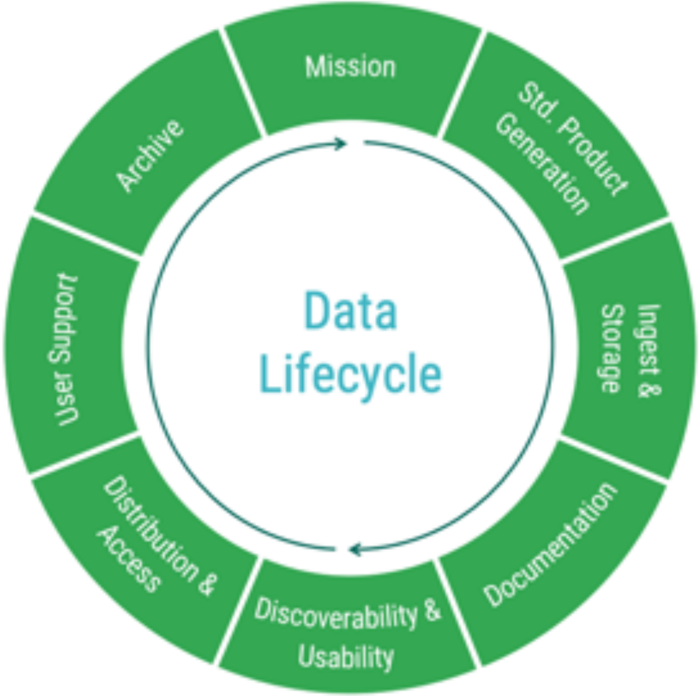
10.04.19

- (PRE) FORMULATION
- IMPLEMENTATON
- PRIMARY OPS
- EXTENDED OPS

NASA's Earth Science Data Systems Program

Single largest repository of Earth Science Data, integrating multivariate/heterogeneous data from diverse observational platforms

Manages NASA's Earth science data through the entire data life cycle



Why is this important?

"The fraction of science papers that rely on archive data is increasing and, in many cases, exceeds the fractions of papers based on new mission data." - NASA Advisory Council Ad-Hoc Big Data Task Force

Enabled by 25+ years of

Open Data
Open Source
Open Services

ESDS by the numbers –FY19

Unique data products
34,500

Data products distributed
1.9 billion

Archived files
462 million

Current archive
33.6 PB

Distinct users
3.5 million

American Customer Satisfaction Index
78

Cumulative Archive Size
Archive Growth Rate
Data Volume (PB)
250
200
150
100
50
15.0
2.6
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Year
NASA

Challenges

- Prepare for planned high-data-rate missions
- Improve efficiency of NASA's data systems operations
- Increase opportunity for researchers and commercial users to access/process PBs of data quickly without the need for data management
- Transparent/extendable open source processing framework
- Ensure users find right data for their problem
- Minimize user burden to access data
- Enable users to extract new knowledge/information from archives

Enabling technologies

Cloud computing

- Operate components of the data systems in a commercial cloud environment to meet future needs
- Provide new opportunities for users to process data in place and perform analytics at scale

Data-driven technologies

- Maximize information and knowledge discovery capabilities
- Augment data stewardship processes
- Address Earth science research and application needs

“Big Data Close to Compute”

Large Volume Data Storage

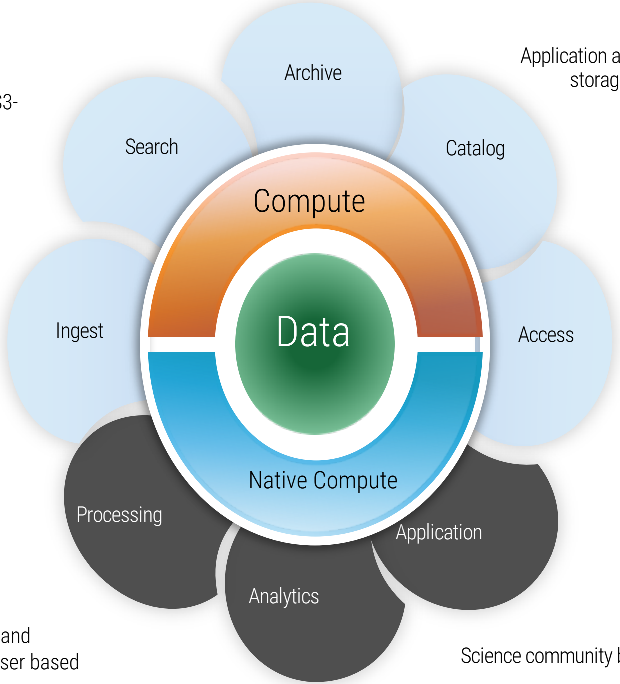
Centralized mission observation & model datasets stored in auto graduated AWS object storage (S3, S3-IA, Glacier)

Scalable Compute

Provision, Access, and terminate dynamically based on need. Cost by use

Cloud Native Compute

Cloud vendor service software stacks and microservices easing deployment of user based applications



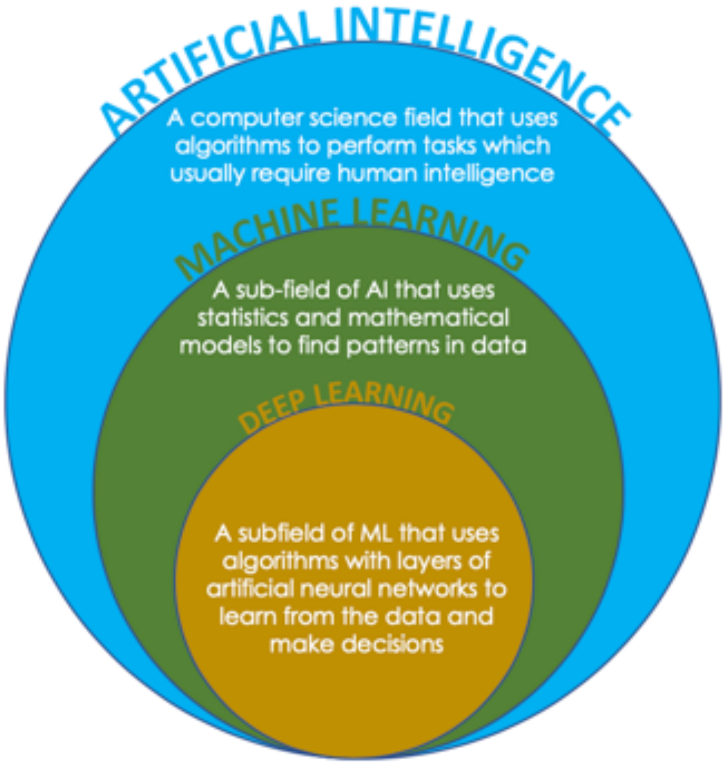
Applications & Services

Application and service layer using AWS compute, storage (S3, S3IA, Glacier), and cloud native technologies

Public Applications & Services

Science community brings algorithms to the data. Support for NASA & non-NASA

Data driven technologies



Rapid adoption of AI/ML due to:

Expanding data volumes

Improving Algorithms

Networks

Cloud computing

Hardware

Maximize information and knowledge discovery capabilities

Phenomena portal

Why?

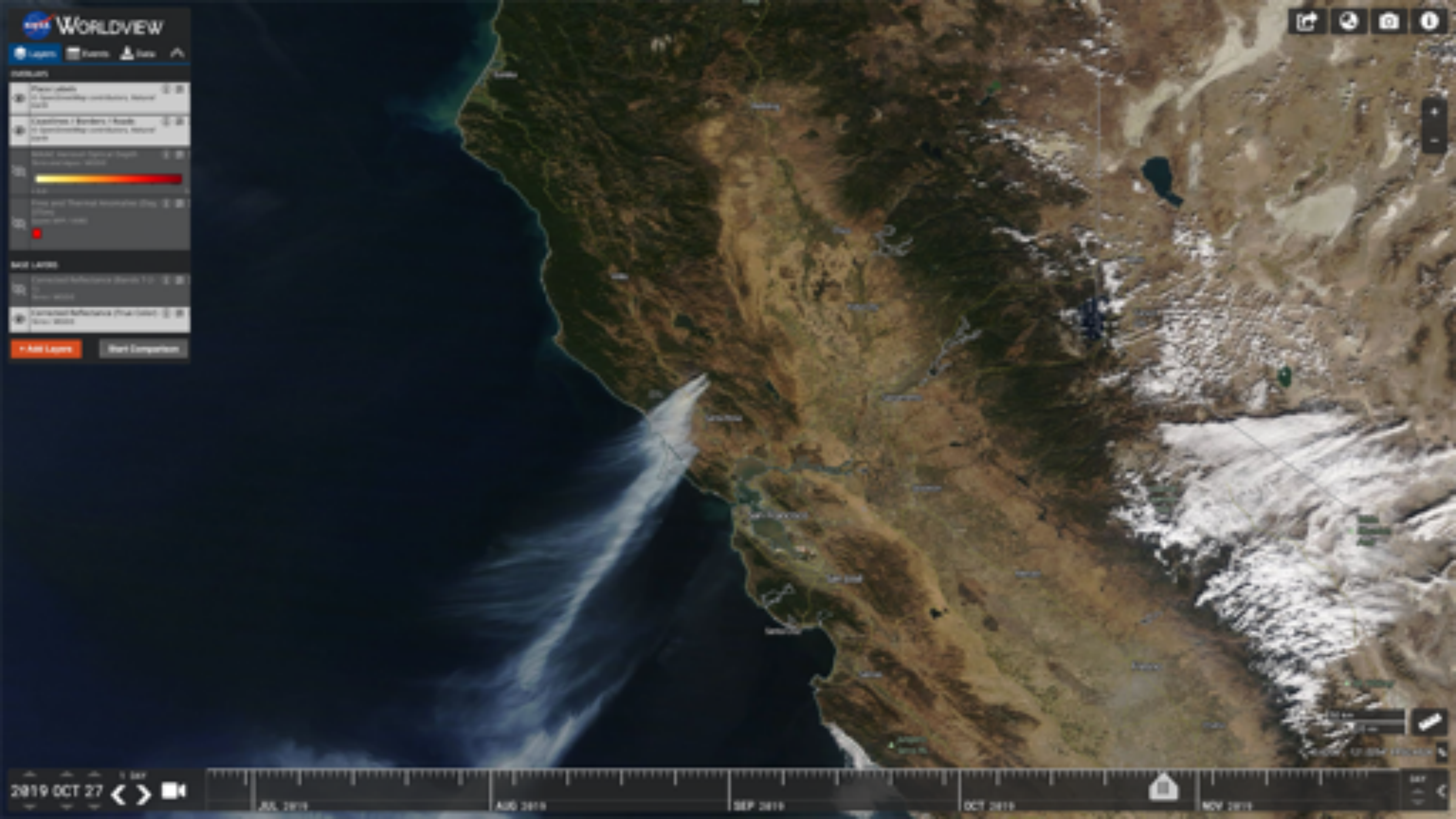
Increasing Earth science data archives require non-traditional approaches to data management

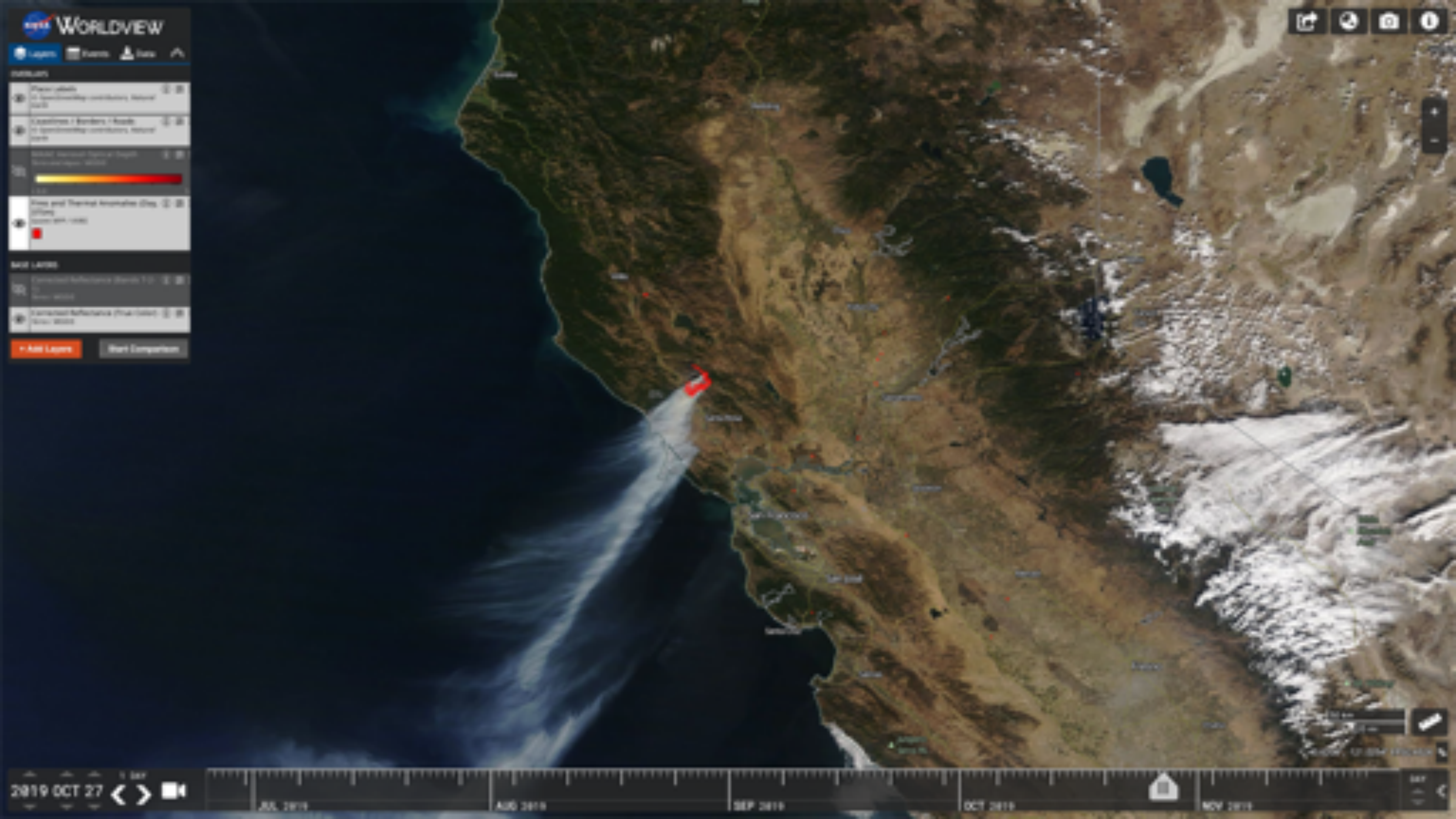
Data driven technologies to provide advanced search capabilities

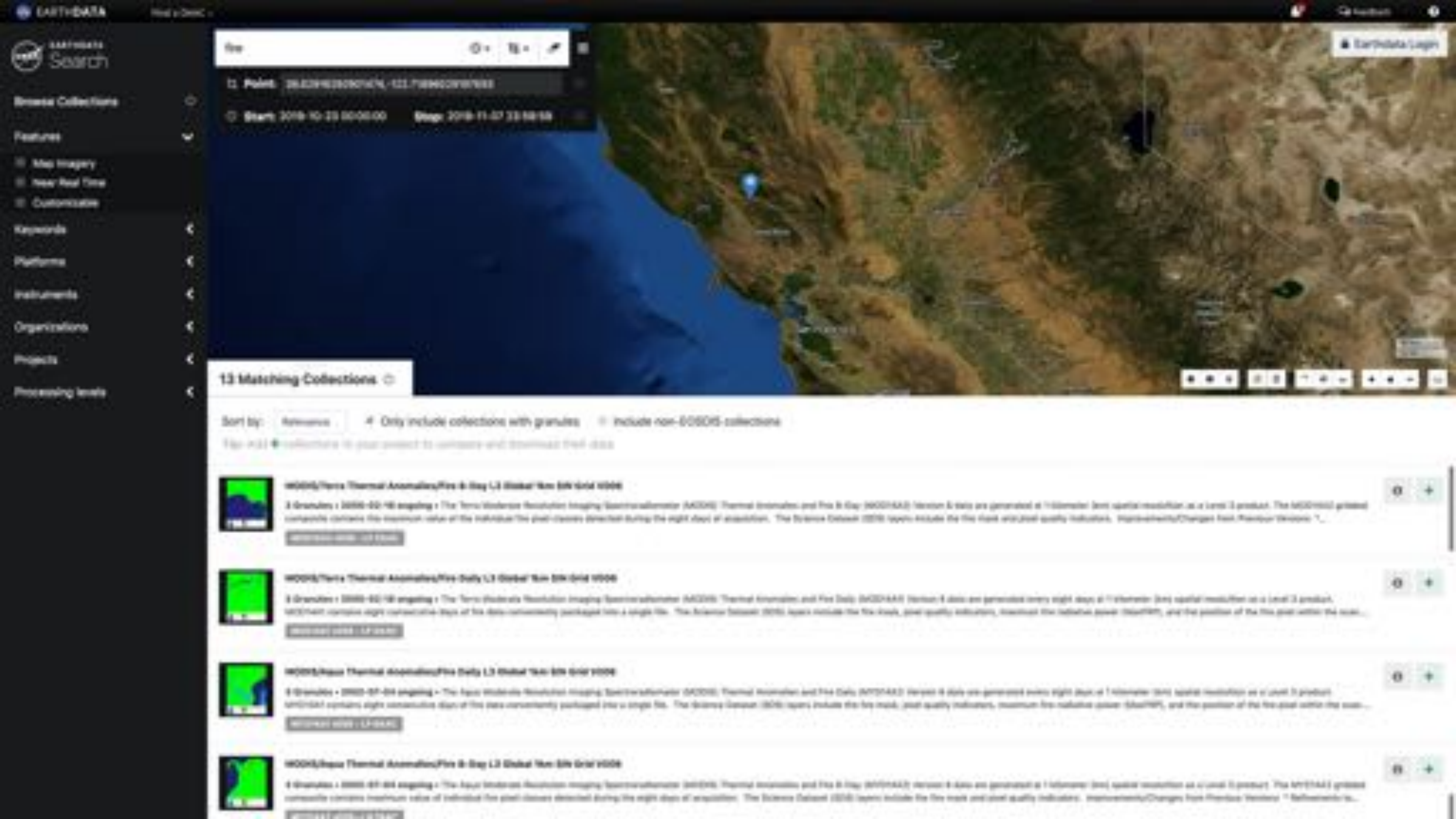
Machine learning-based approach - an enabling data driven technology to provide automated detection of Earth science events from image archives

Catalog of events can provide a novel way to explore large archives of data

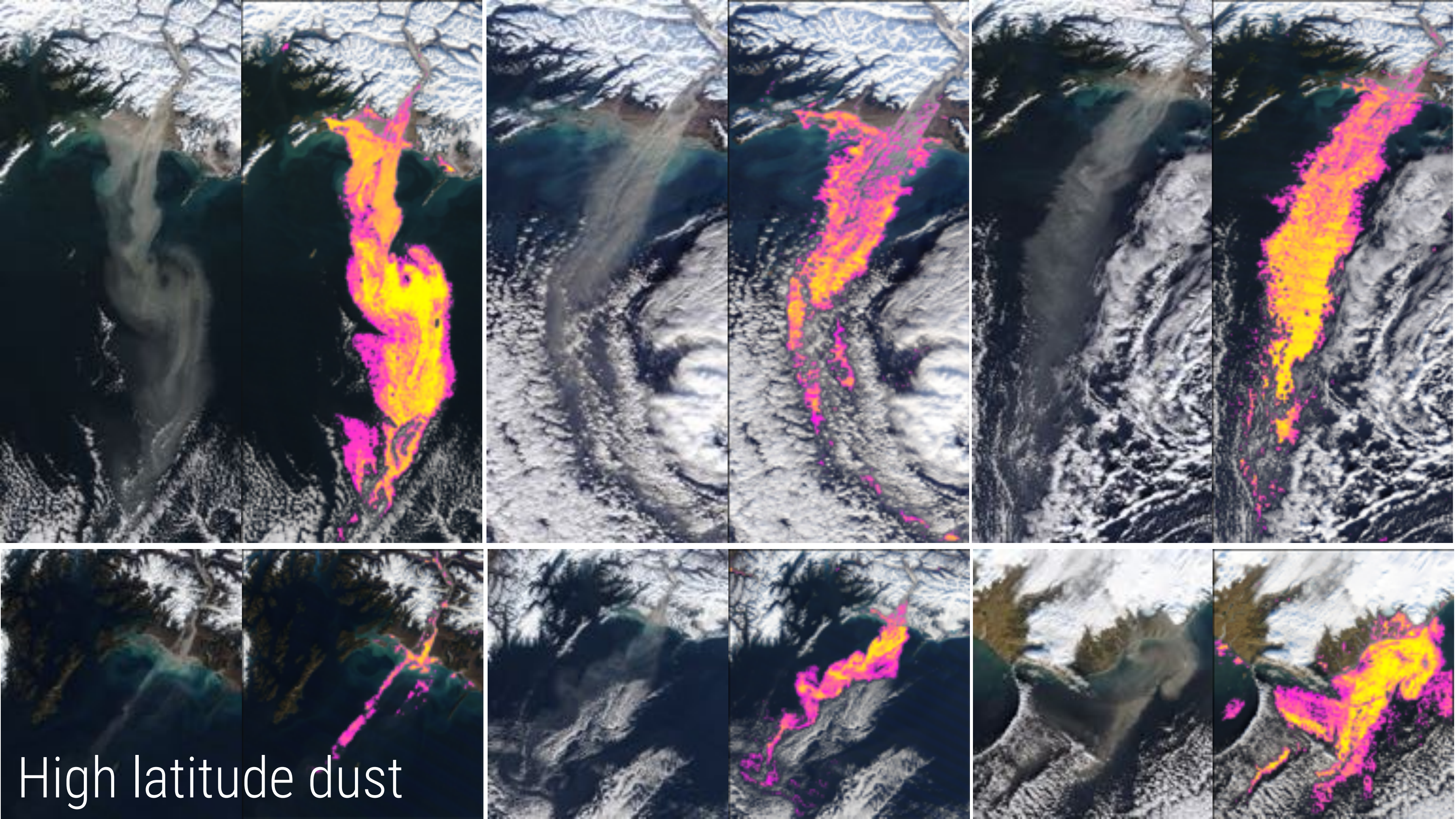
Discover and explore Earth science data archives around events using machine learning (ML) techniques


WORLDVIEW
+ Add Layers
Start Comparison
2019 OCT 27
JUL 2019
AUG 2019
SEP 2019
OCT 2019
NOV 2019

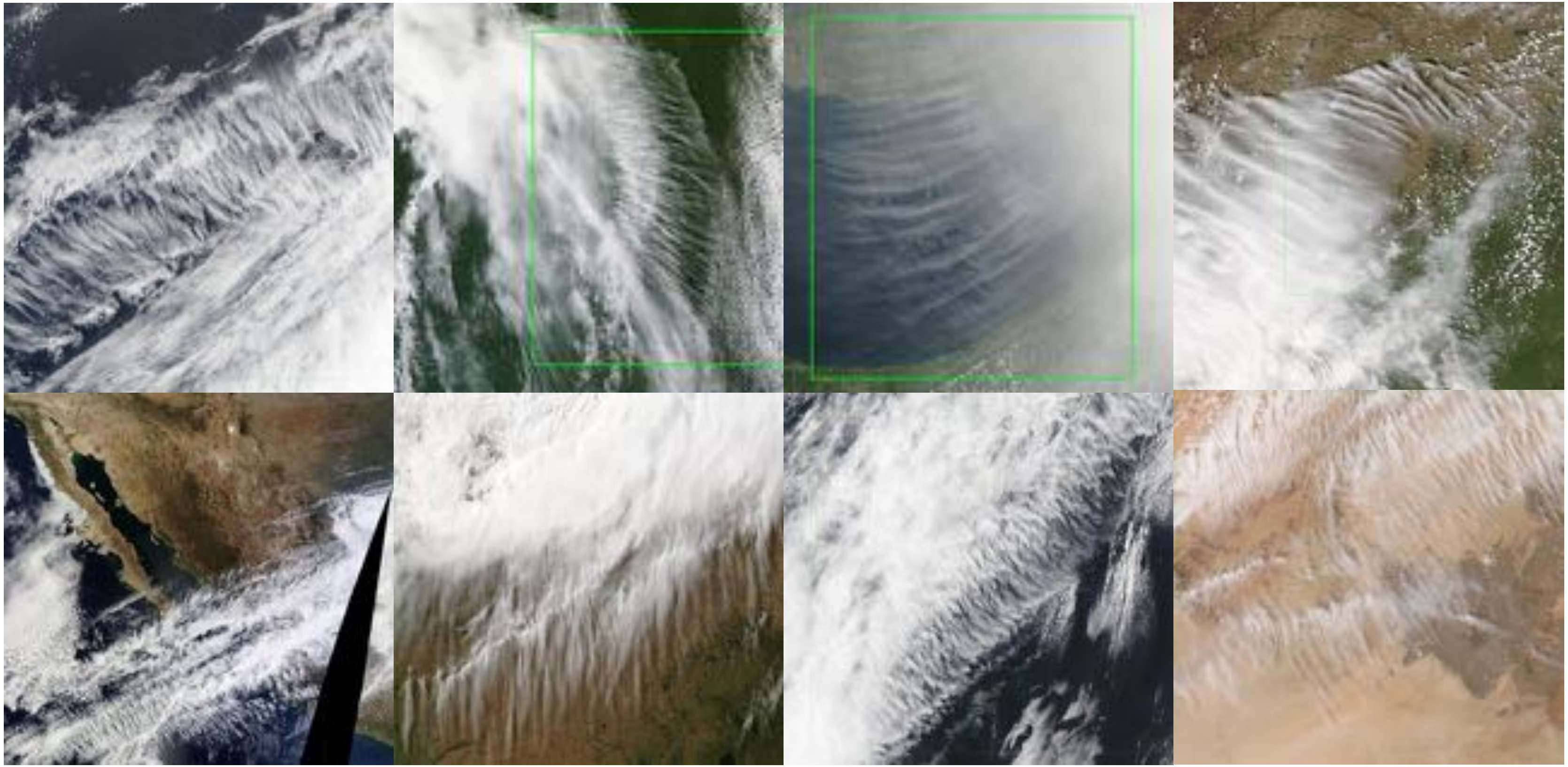

WORLDVIEW
+ Add Layers
Start Comparison
2019 OCT 27
JUL 2019
AUG 2019
SEP 2019
OCT 2019
NOV 2019


EARTHDATA
Earthdata Search
Earthdata Login
fire
Point:
Start:
Stop:
Browse Collections
Features
Map Imagery
Near Real Time
Customizable
Keywords
Platforms
Instruments
Organizations
Projects
Processing levels
13 Matching Collections
Sort by:
Relevance
Only include collections with granules
Include non-EOSDIS collections
MODIS/Terra Thermal Anomalies/Fire 8-Day L3 Global 1km SIN Grid V006
MODIS/Terra Thermal Anomalies/Fire Daily L3 Global 1km SIN Grid V006
MODIS/Aqua Thermal Anomalies/Fire Daily L3 Global 1km SIN Grid V006
MODIS/Aqua Thermal Anomalies/Fire 8-Day L3 Global 1km SIN Grid V006

Kincade Fire Demo Video


High latitude dust

Transverse cirrus bands



Phenomena Portal Demo Video

Augment data stewardship processes

Automated keyword assignment

Why?

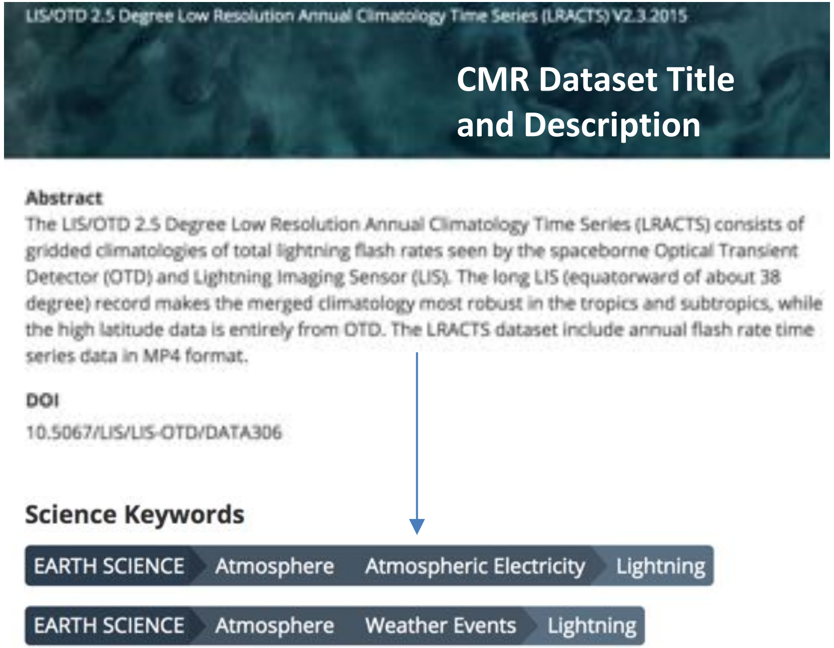
Assigning science keywords is currently a manual process, which is prone to human error and inconsistencies.

Metadata managed across a network of multiple data centers (i.e. keywords not assigned by a central entity)

Keywords may be assigned by non-subject matter experts (SMEs)

Improve metadata quality

Provide objective and consistent approach to keyword assignment



Approach – build word embeddings

88,410

documents

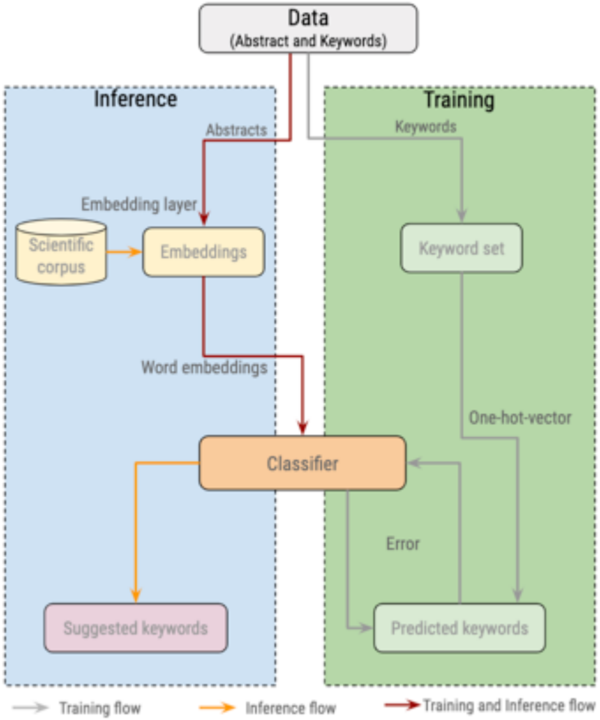
530 million

words

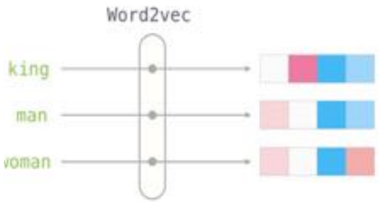
5.5 million

unique words

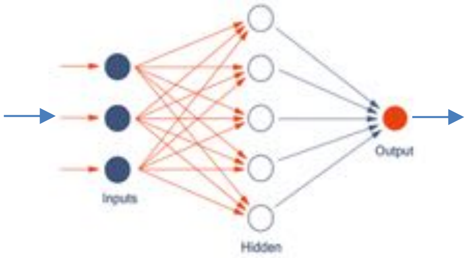
Automated keyword assignment



Version 7.3 is the current version of the data set. Version 3.5 is no longer available and has been superseded by Version 7.3. This data set is currently provided by the OCO (Orbiting Carbon Observatory) Project. In expectation of the OCO-2 launch, the algorithm was developed by the Atmospheric CO2 Observations from Space (ACOS) Task as a preparatory project, using GOSAT TANSO-FTS spectra. After the OCO-2 launch, "ACOS" data are still produced and improved, using approaches applied to the OCO-2 spectra. The "ACOS" data set contains Carbon Dioxide (CO2) column averaged dry air mole fraction for all soundings for which retrieval was attempted. These are the highest-level products made available by the OCO Project, using TANSO-FTS spectral radiances, and algorithm build version 7.3. The GOSAT team at JAXA produces GOSAT TANSO-FTS Level 1B



word2vec



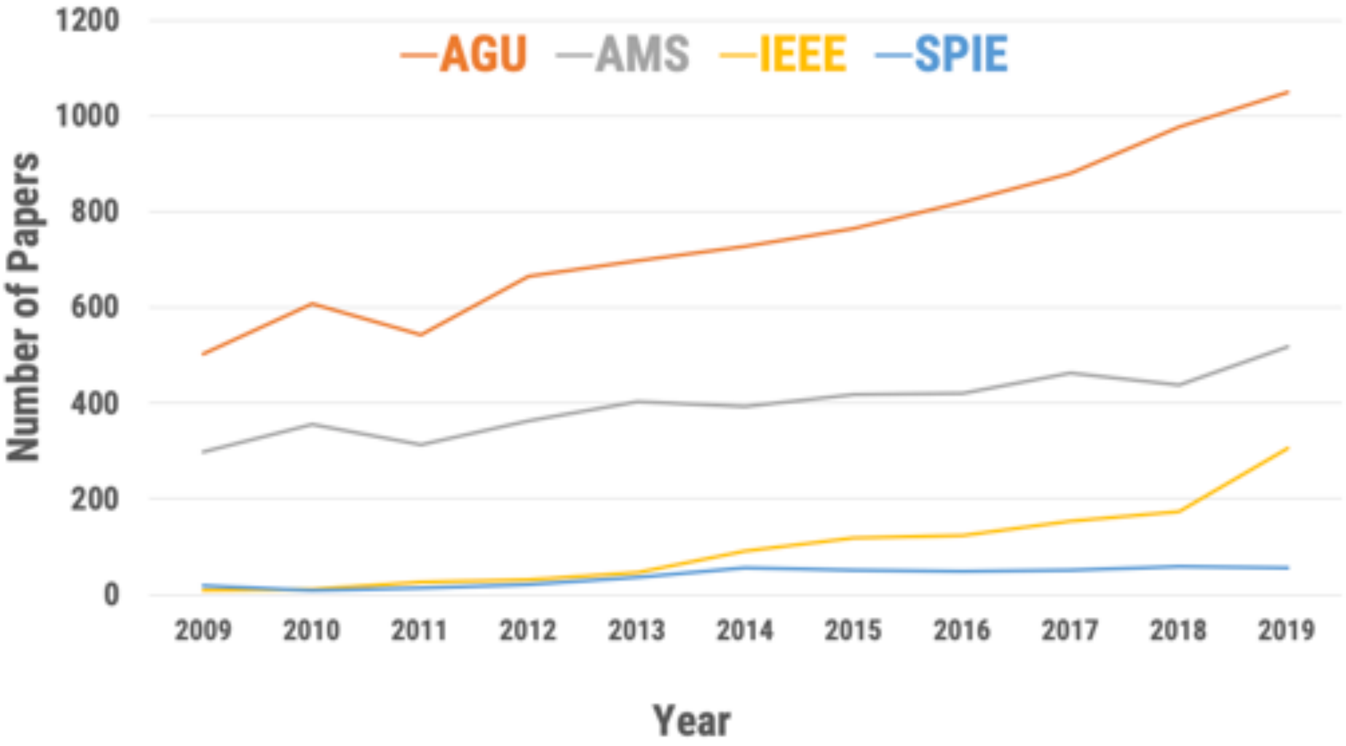
classifier

Predicted Keyword	Score
carbon dioxide	0.4513424
land use/land cover classification	0.3825603
terrain elevation	0.1924277
barometric altitude	0.18085223
carbon and hydrocarbon compounds	0.07634798

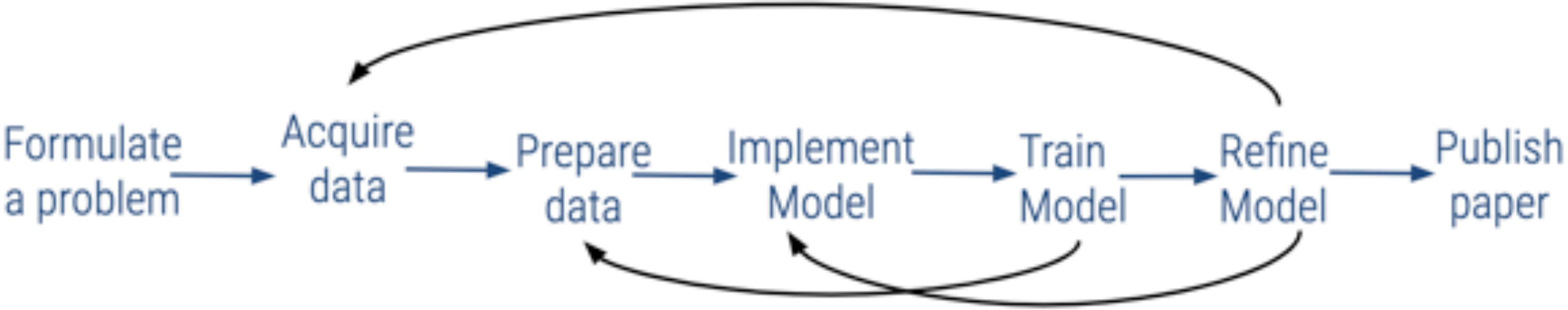
Address Earth science research and application needs:

Hurricane intensity estimation system

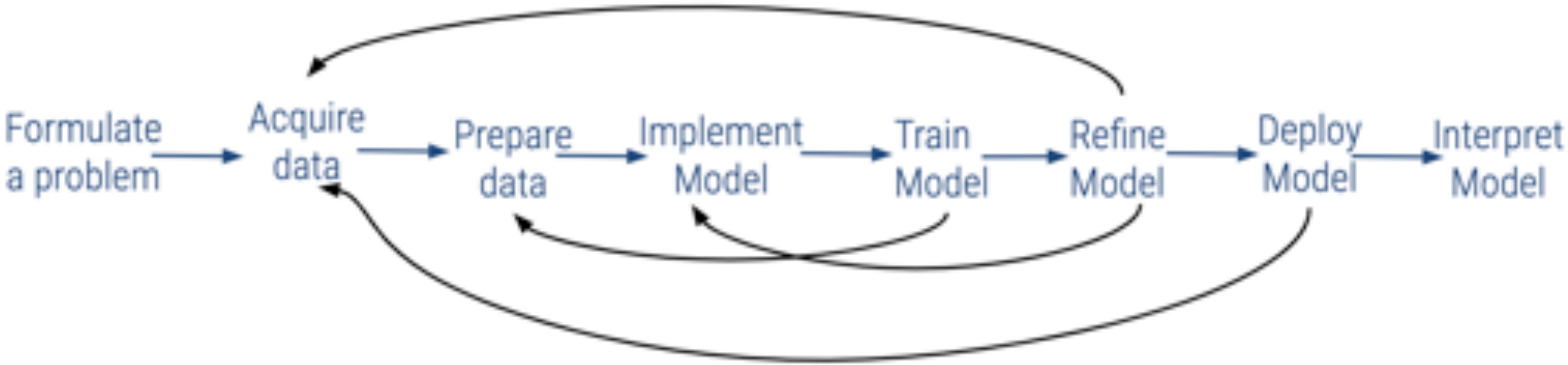
AI/ML in Earth Science



ML in literature



ML lifecycle - iterative



Motivation

15 UTC 10 Oct 17 NHC advisory on Tropical Storm Ophelia:

"Dvorak intensity estimates range from T2.3/33 kt from UW-CIMSS to T3.0/45 kt from TAFB to T4.0/65 kt from SAB. For now, the initial intensity will remain at 45 kt, which is an average of the scatterometer winds and all of the other available intensity estimates."

Motivation

15 UTC 10 Oct 17 NHC advisory on Tropical Storm Ophelia:

"Dvorak intensity estimates range from T2.3/33 kt from UW-CIMSS to T3.0/45 kt from TAFB to T4.0/65 kt from SAB. For now, the initial intensity will remain at 45 kt, which is an average of the scatterometer winds and all of the other available intensity estimates."

Motivation

15 UTC 10 Oct 17 NHC advisory on Tropical Storm Ophelia:

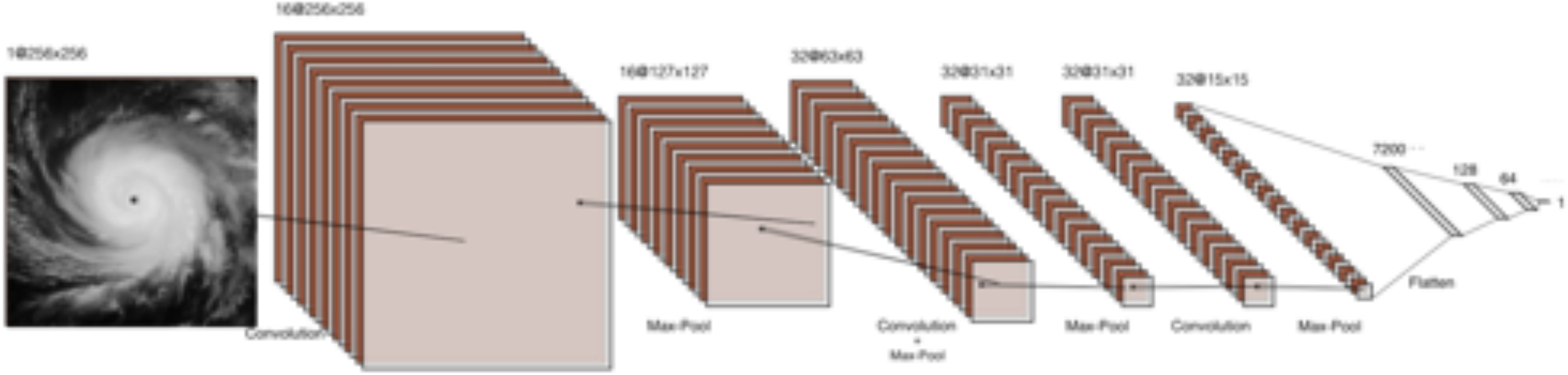
"Dvorak intensity estimates range from T2.3/33 kt from UW-CIMSS to T3.0/45 kt from TAFB to T4.0/65 kt from SAB. For now, the initial intensity will remain at 45 kt, which is an average of the scatterometer winds and all of the other available intensity estimates."

Can we objectively estimate wind speed from satellite images?

Can we estimate more frequently?

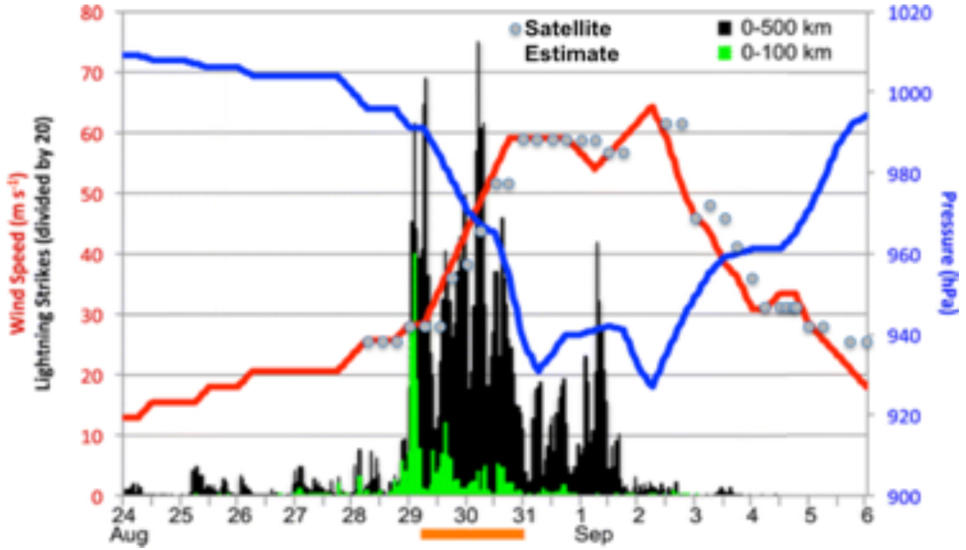
Data

Model development



Test results

Detailed look: Hurricane Earl, 2010



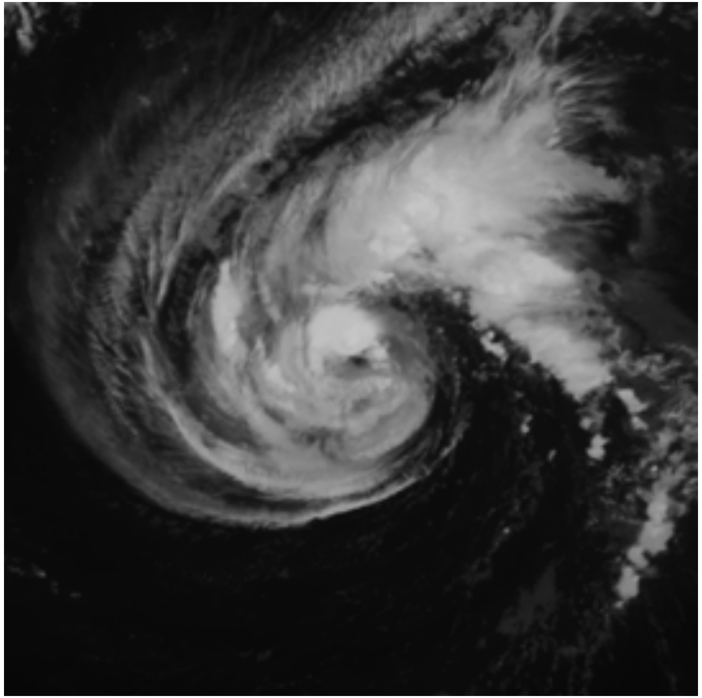
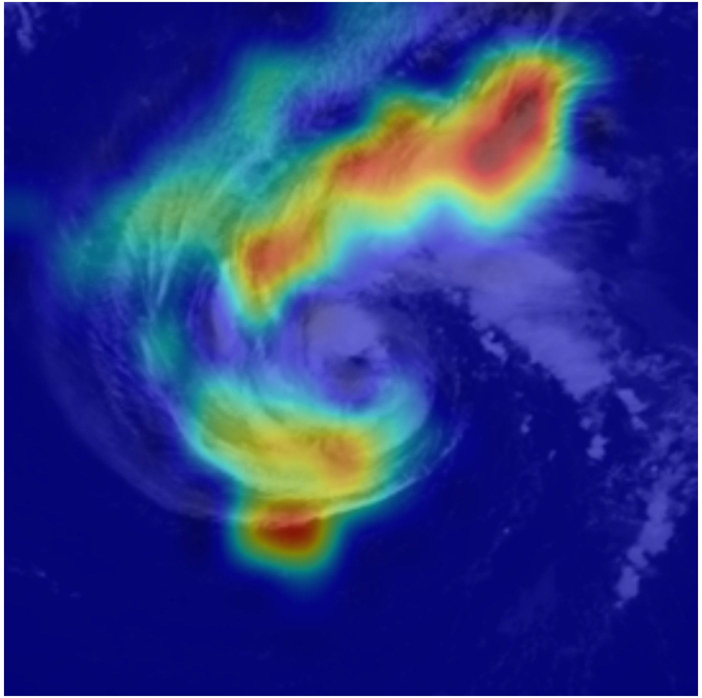
Adapted from Stevenson et al. (2014). Time series of satellite-derived intensity estimates (circles) for Hurricane Earl (2010), added to best track intensities and lightning flash rate time series.

AI as a black box

Interpretability + model inspection

Learning Deep Features for Discriminative Localization

Model evaluation with class activation maps

Tropical Storm

Hurricane Dorian Demo Video

We have a model…..now what?

Going extra mile

Interpret prediction data – prediction output maybe just numbers

Questions:

Does the model confidence remain the same over time?
How do you maintain?
How do you complete the loop with new training data?

Deployment to production

Performance requirements

Metrics and baselines with initial models

Monitor over time

Back-testing

Model and software will change

Testing model changes on historical data

Run current production model to baseline performance

Run new models, competing for production

Now-testing

Testing of production model on latest data

Can we get early warning that the model may be faltering?

- Content drift: training data exploited by model are subtly changing with time

Hurricane wind speed estimation portal

Features of a situational awareness tool:

Monitor NHC outlook for “invest” area for trigger

Near real-time tropical cyclone intensity estimation services

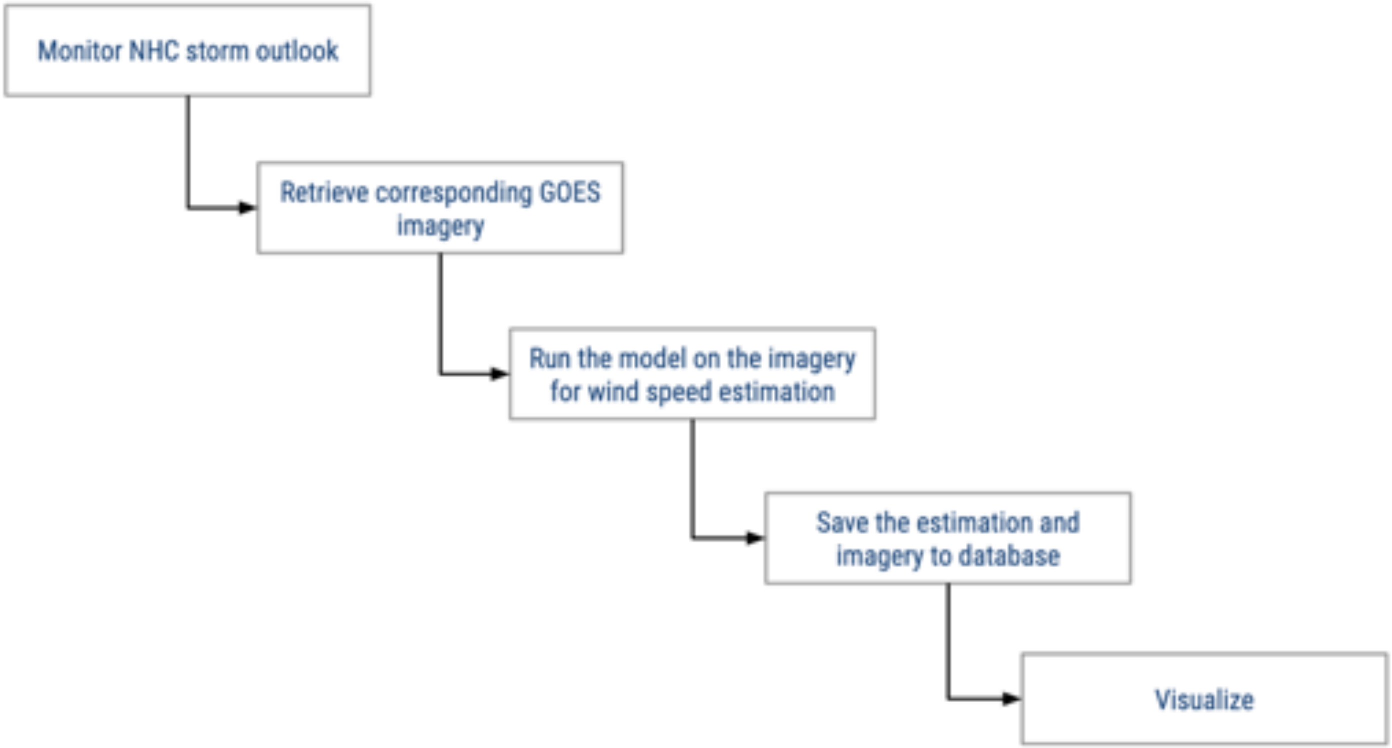
Map display

Layers

Comparison with operational forecasts/Evaluate

Service APIs

Workflow


Monitor NHC storm outlook
Retrieve corresponding GOES imagery
Run the model on the imagery for wind speed estimation
Save the estimation and imagery to database
Visualize

Coordinated effort

ML researchers

- Transform ideas into models
- Training data
- Monitor

Domain experts

- Evaluation
- Performance baselines
- Science use case

End-user stakeholders

- Production requirements

ML engineers

- Design
- Quick prototype
- Deployment
- Scale
- Log

Hurricane Intensity Estimation Portal Demo Video

Challenges and lessons learned

Consistent large-scale training data

AI black box

Training data/Input data becomes part of the code

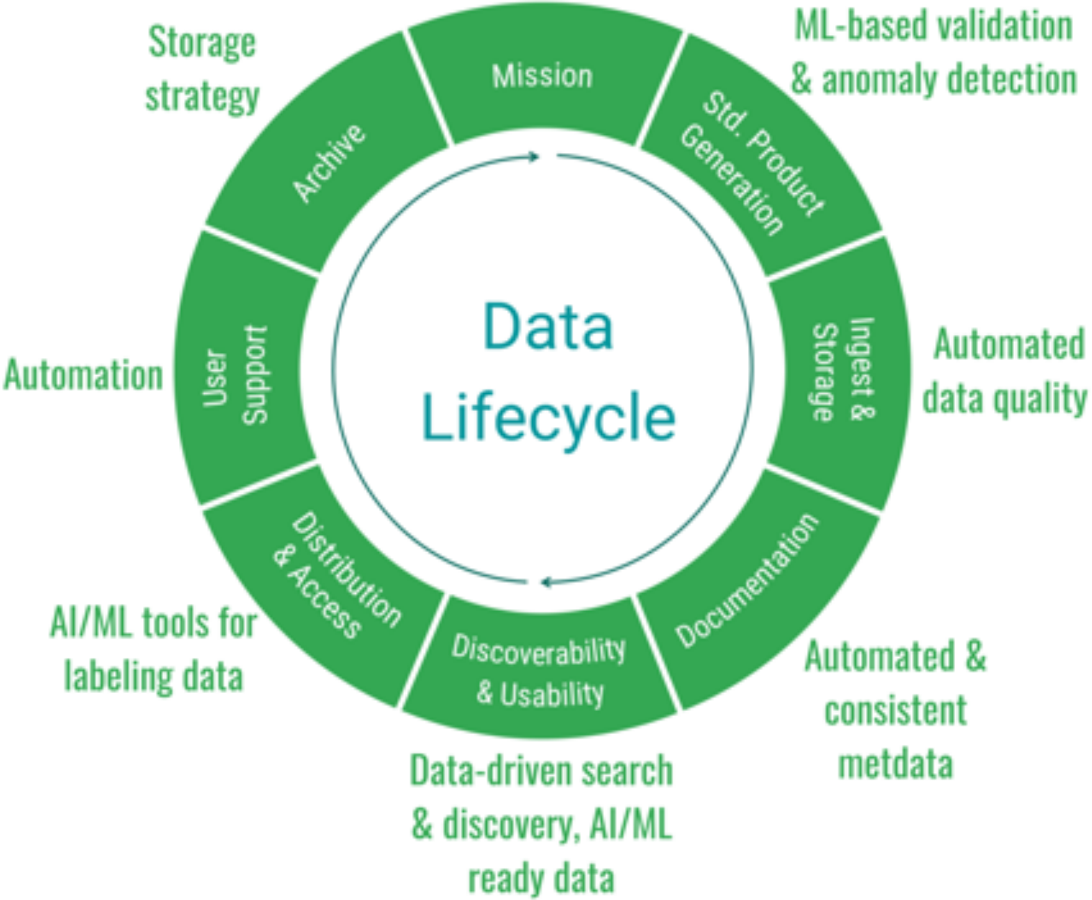
Versioning training data, model, algorithm becomes difficult

DevOps, CI/CD

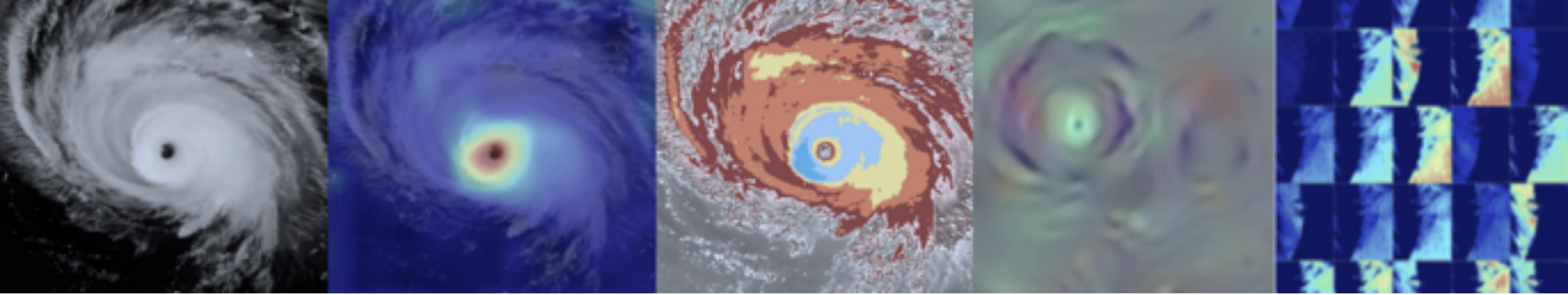
Complexity with evolving platforms and infrastructure

What’s next?

Biggest bottleneck in adoption of ML in Earth Science is training data



AI for full data lifecycle



Thank you.

Manil Maskey
manil.maskey@nasa.gov